ChinaXiv合作期刊

干旱区地理
ARID LAND GEOGRAPHY

基于主成分和机器学习的土壤有机质含量空间预测建模

胡贵贵[1,2]，杨粉莉[3]，杨联安[1,2]，郑玉蓉[1,2]，王　辉[4]，陈卫军[5]，李亚丽[1,2]

（1. 西北大学陕西省地表系统与环境承载力重点实验室，陕西 西安 710127；2. 西北大学城市与环境学院，陕西 西安 710127；3. 咸阳市农业科学研究院，陕西 咸阳 712000；4. 咸阳市土壤肥料工作站，陕西 咸阳 712000；5. 旬邑县土壤肥料工作站，陕西 旬邑 711300）

摘　要：协同环境变量与机器学习回归模型构建土壤有机质空间预测组合模型对养分精准管理具有重要意义，而多维变量间的信息冗余和相关性会导致模型训练时间过长、预测精度降低等问题。以陕西省咸阳市农耕区为例，选取高程、坡向、坡度、剖面曲率、平面曲率、地形起伏度、地形湿度指数、年均降水量、年均气温、归一化植被指数共10个环境变量，在主成分分析（Principal component analysis，PCA）、核主成分分析（Kernel principal component analysis，KPCA）方法特征提取基础上，组合随机森林（Random forest，RF）、支持向量回归机（Support vector regression，SVR）、K最近邻（K-nearest neighbor，KNN）机器学习模型进行土壤有机质含量空间预测。以单一模型作为对照，通过计算模型决定系数（Coefficient of determination，R^2）、均方根误差（Root mean square error，RMSE）和相对绝对误差（Relative absolute error，RAE），对不同模型的预测结果进行精度评价。结果表明：利用主成分提取方法和机器学习算法构建组合模型能消除变量间相关性，一定程度上提高土壤有机质含量预测模型精度。KPCA-RF模型对SOM含量预测精度高于其他模型，R^2、RMSE、RAE分别为0.791、1.970 g·kg^{-1}、50.100%，该模型良好的预测能力可以为土壤有机质含量的空间预测与制图提供科学依据。

关 键 词：土壤有机质；机器学习；核主成分分析；农耕区；咸阳市

文章编号：

chinaXiv:202108.00019v1

土壤有机质（Soil organic matter，SOM）是衡量土壤肥力与土壤质量的重要指标之一，对维持土壤生态平衡与促进作物生长发育具有至关重要的作用[1]。传统的土壤制图方法依赖于大量的采样点数据来表达土壤养分的空间分布特征，这种方式耗费大量的人力物力，时间周期长且精度难以保证。随着新的研究方法、技术手段以及认知水平的提高，数字土壤制图成为一种高效表达土壤空间分布的新方法[2-3]，可以获取精细土壤信息以便于指导农业生产与田间管理。

数字土壤制图以土壤-景观模型为理论基础，借助空间分析和数学方法等技术手段进行土壤调查和可视化的现代化技术体系[2]。通过选取与土壤发生相关和体现土壤属性空间差异的表征因子作为环境变量，分析不同变量条件下土壤属性的差异性并建立它们之间的逻辑关系，从而推测出土壤属性的空间连续分布特征[3]。自20世纪90年代Moore等[4]开始采用线性回归与判别分析对土壤属性进行预测，学者们对于数字土壤制图模型方法的探索从未停止，大致经历了线性回归、模糊逻辑推理、地统计学、机器学习等过程[5]。线性回归模型依靠已知点的变量值与辅助信息建立简单线性、多元线性回归方程，从而实现其余未知点的预测[6-7]，但是未考虑到样点间的相关性，预测偏差较大；模糊逻辑推理方法将土壤与环境条件知识表达为隶属度函数，某个未知点与多个采样点土壤间具有隶属度，根据

收稿日期：2020-06-09；　**修订日期：**2020-12-21

基金项目：国家自然科学基金（41771129）；陕西省农业科技攻关项目（2011K02-11）资助

作者简介：胡贵贵（1996-），男，硕士研究生，主要从事RS与GIS在农田土壤养分中的应用研究. E-mail: 17634976916@163.com

通讯作者：杨联安（1968-），男，博士，副教授，主要从事地理信息系统及土壤肥料学研究. E-mail: yanglianan@163.com

ChinaXiv合作期刊

这些隶属度确定未知点的土壤属性，以 Zhu 等[8]提出的 SoLIM 模型作为代表；地统计学模型以区域化变量理论为基础，通过已知点的空间依赖性预测未知点的变量值[9]，如普通克里格。协同克里格结合了空间相关理论、要素间土壤属性和环境变量间协同相关性对目标属性进行预测估计，一定程度上能提高土壤推测的精度[7,10]；机器学习在处理多维、非线性海量数据、改善模型泛化能力等方面具有良好的适用性，目前已经被应用到诸多领域的研究中[11-12]。尤其在展示土壤空间变异、土壤养分空间预测方面，协同多源环境变量的机器学习方法显示出较大的潜力，常见的模型包括随机森林（Random forest，RF）[13-14]、支持向量机（Support vector machine，SVM）[15]、决策树（Decision tree，DT）[16]、神经网络（Neural networks，NNs）[17]、K-最近邻（K-nearest neighbor，KNN）[18]等。Curtis 等[19]使用逐步线性回归、决策树、RF 等 8 种统计与机器学习算法对玉米种植区土壤氮需求进行预测，结果发现机器学习模型的预测精度更高；Tomislav 等[16]研究显示 RF 模型对土壤属性的预测误差较线性回归模型下降了 15%~75%，表明 RF 模型在非洲土壤属性制图中同样具有良好的适用性；类似的研究[20-21]也表明机器学习模型较传统线性模型能更好地刻画土壤属性的空间变异性。

机器学习模型在数字土壤制图应用中起步较晚，不同区域下模型的适用性和预测性能仍需进一步研究，其中环境变量的选取也成为研究的重点[22]。土壤有机质空间分布受地形、气候、植被和人为活动等多方面的影响，通过探寻土壤与环境之间的关系，不断丰富土壤环境关系库对数字土壤制图具有重要的意义[23]。地形是影响土壤发育的重要因素之一，通过调节土壤水分与太阳辐射的空间再分配对土壤养分产生影响，同时数字高程模型（Digital elevation model，DEM）提供了高分辨率的地表状况且易获取，在土壤预测制图中得到了广泛的应用[13]；遥感影像提供了完整的、高精度的地表反射率数据，可通过反演相应的盐分指数[23]、植被指数[14]来表征土壤属性地表差异；降水量和气温等气候因子主要是通过干预植物生长发育与有机质分解影响土壤有机质的空间分布。此外，人为活动因素由于数据的获取难度较大、目前缺乏有效的方法进行空间量化而较少参与土壤养分预测制图。

随着越来越多的变量参与到机器学习模型构建中，多维变量间的信息冗余和相关性导致模型训练时间加长、预测精度出现偏差等问题。利用主成分提取方法与机器学习算法构建组合模型是 1 种可行的优化策略[24-25]，主成分提取方法通常分为线性和非线性 2 种，线性方法有 PCA（Principal component analysis）、典型相关分析、线性判别分析等，非线性方法有 KPCA（Kernel principal component analysis）、流形学习等。学者们的研究大多集中于线性提取方法，尤其以 PCA 方法最多[25-26]，组合非线性方法和机器学习模型的研究相对较少。本研究选取地形、气候、植被 3 大类共 10 个环境变量作为输入变量，在 PCA、KPCA 2 种主成分特征提取的基础上，结合 RF、SVR、KNN 3 个模型，构建相应的组合模型，为咸阳市农耕区 SOM 含量的空间模拟预测提供科学依据。

1 研究区概况及数据来源

1.1 研究区概况

陕西省咸阳市位于 107°38′~109°10′E，34°11′~35°32′N 之间，地处陕西关中平原腹地，东西跨度约 139.7 km，南北跨度约 149.4 km，总面积约为 10189.4 km^2。咸阳市属于暖温带大陆性季风气候，全年平均降水量为 537~650 mm，平均温度 9.0~13.2 ℃。根据地形变化，全市可分为 3 个大的地貌分区：南部关中平原区、北部黄土高原区和东北部山地区。受地形条件的影响，南北地区热量条件呈现明显的差异，年均气温南部一般比北部高 4.2 ℃，北部无霜期为 172~205 d，南部无霜期为 212~223 d。境内农耕区主要分布在平原和高原区域，由于东北部山区耕地较少，故未将该区域纳入本次的研究范围。

1.2 数据来源

1.2.1 *土壤数据的采集* 参照研究区土壤特性、地貌特点、作物信息对农耕区（不包括秦都区与渭城区）进行样点布置，结合农业部测土配方施肥技术规范，遵循均匀性、代表性、多点混合的原则进行采样。采样时间为 2017 年作物收获后，采用“S”形法均匀随机取 5 个点，将各采样点土壤混匀后用四分法留取 1 kg 土样装袋，采样深度为 0~20 cm，同时利用 GPS 记录样点的经纬度位置，登记土样编号，记

chinaXiv:202108.00019v1

ChinaXiv合作期刊

录土样的种植制度、灌溉条件、产量等相关属性。在实验室经风干、研磨和过筛后，采用重铬酸钾氧化容量法测定SOM含量，并选用域值法(3倍标准差)剔除异常值，最终得到407份有效样点数据，样点分布如图1所示。

1.2.2 环境变量收集与处理 环境变量包括地形因子、气候因子、植被因子。地形因子源于从地理空间数据云平台下载的90 m分辨率数字高程模型数据，利用ArcGIS 10.3提取出研究区高程(DEM,*X*1)、坡度(Slope,*X*2)、坡向(Aspect,*X*3)、剖面曲率(*X*4)、平面曲率(*X*5)、地形起伏度(Relief degree of land surface, RDLS, *X*6)和地形湿度指数(Topographic wetness index,TWI,*X*7)7种地形因子，其中RDLS和TWI的计算公式分别参见文献[27-28]；从世界气象数据库下载全球多年(1970—2000年)月平均降水量(1—12月)和月平均气温(1—12月)数据，空间分辨率为1 km。通过栅格计算、掩膜提取得到研究区年均降水量(*X*8)和年均气温(*X*9)数据，最后重采样将空间分辨率转换为90 m；植被因子采用归一化植被指数(NDVI,*X*10)，用来反映农田植被生长状况，使用2017年7月的Landsat 8遥感影像反演提取所得。

1.2.3 构建预测因子集合 根据采样点的空间位置信息，利用ArcGIS 10.3软件Spatial analyst模块中提取分析将10个环境变量值提取至每一样点，构建土壤有机质预测因子集合。

2 方法与模型

2.1 主成分提取方法

在土壤养分空间预测研究中，加入不同的环境变量到预测模型中能够有效地提高模型精度。但不同变量间往往存在着较高的相关性，大量的冗余信息会导致模型精度出现偏差。主成分提取就是将原有变量的有用信息集中在尽可能少的新的主成分变量中，达到信息增强的目的。

PCA作为1种常见的多元统计方法，通过探索变量间的线性关系，从而实现将多维变量综合成少数变量的线性组合，是进行数据降维、模型优化的1种有效手段。KPCA是1种非线性的特征提取方法，它的主要思想是：通过非线性核函数转换的方法，将原始向量映射到高维特征空间中，然后在特征空间中进行线性主成分变换[29]，常见的核函数有：线性核函数、多项式核函数、高斯核函数。KPCA克服

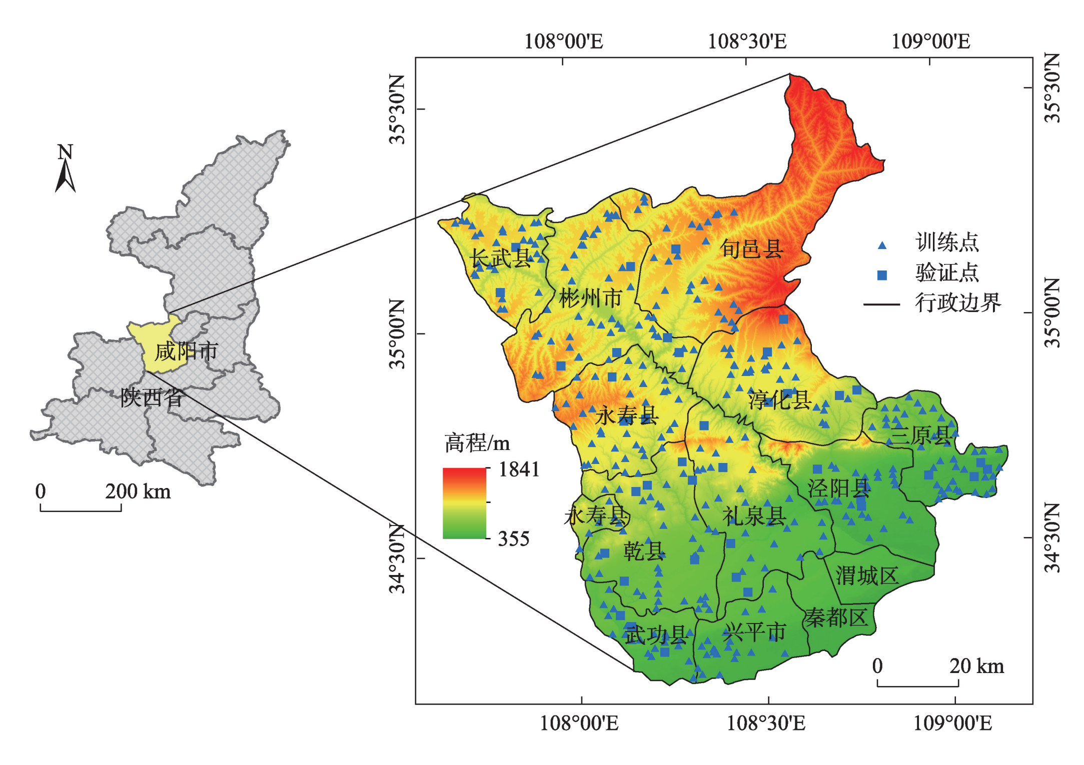


图1 土壤有机质(SOM)采样点分布图

Fig. 1 Distribution map of soil organic matter sampling points

chinaXiv:202108.00019v1

ChinaXiv合作期刊

了PCA方法不能反映数据的非线性特征以及只考虑二阶统计特性的缺陷，在图像降维[30]、生态评价[31]等方面得到广泛的应用。

2.2 机器学习模型简介

2.2.1 随机森林 随机森林（RF）是Breiman[32]提出的基于决策树的分类回归算法，它是利用bootstrap重抽样方法从原始数据集中随机抽取多个样本，对每个样本数据进行决策树建模，然后组合成多棵决策树的预测模型，综合所有决策树的投票结果得到模型的预测值，具体算法步骤见文献[33]。

2.2.2 支持向量回归机 支持向量回归机（Support vector regression，SVR）是基于VC维理论和结构风险最小化原则，从已知的样本数据中获取最佳的学习模型[34]。它使用"核技巧"方法，将低维特征空间映射到一个高维甚至无穷维的特征空间中，从而使用线性回归的方法实现低维特征空间数据的回归预测。

2.2.3 K最近邻算法 K最近邻算法（KNN）依赖于欧几里得距离，通过距离排序选择K个最近的样本来预测目标[35]。

2.3 研究流程

主成分提取、模型的训练与测试分别在Matlab、Rstudio软件中进行，主要步骤包括：

（1）数据预处理。为了消除量纲，对构建的SOM预测因子集合数据进行归一化处理，将所有变量值域放在相等的区域[36]。

（2）特征提取。运用PCA、KPCA分别进行数据处理，选择累计贡献率大于85%的主成分作为后续模型的输入特征数据。

（3）划分训练集、验证集。通过随机抽样的方法，按照7:3的比例划分训练集、验证集。

（4）训练模型。根据285个训练样本，分别调用randomForest、e1071、caret语言包构建基于PCA的PCA-RF、PCA-SVR、PCA-KNN预测模型、基于KPCA的KPCA-RF、KPCA-SVR、KPCA-KNN预测模型。进一步对各模型参数进行遍历，根据10折交叉验证结果选取参数最优值。

（5）精度评价。采用验证集测试SOM含量预测模型性能，与未进行主成分分析的RF、SVR、KNN单一预测模型对比分析，实现对咸阳市农耕区SOM含量的精准预测。本文选择模型决定系数（Coefficient of determination，R^2）、均方根误差（Root mean square error，RMSE）和相对绝对误差（Relative absolute error，RAE）作为土壤有机质预测模型的评价指标。

3 结果与分析

3.1 描述性统计

由表1可知，咸阳市SOM含量介于6.59~27.80 g·kg^{-1}，平均值为15.54 g·kg^{-1}，处于《全国第二次土壤普查养分分级标准》的第4级即10~20 g·kg^{-1}，处于稍微缺乏状态；训练集、验证集和整体数据集除了样点数不同，其他各统计参量相差不大；变异系数接近25.00%，属于中等变异性。

3.2 基于主成分和机器学习的SOM含量预测建模

使用主成分方法进行数据特征提取，在保留大部分原始信息的同时，可以降低变量间的相关性。基于407个样点数据，分析SOM含量与各环境变量的相关性、不同环境变量之间的相关程度。由表2可知：SOM含量与$X1$、$X6$呈显著的负相关关系（$P<0.01$），与$X8$、$X9$、$X10$呈显著的正相关关系（$P<0.01$），与$X2$、$X7$两因子相关性在0.05置信水平上呈显著相关。其中，SOM含量与DEM之间相关性最强，相关系数为-0.315，表明SOM含量随海拔的升高而呈下降趋势，不同高度的SOM含量差异明显。同时结果也显示：各环境变量之间存在相关性，变量$X1$与$X2$、$X6$、$X7$、$X8$、$X9$、$X10$之间呈现显著相关（$P<0.01$），与变量$X3$之间有着较强的相关性（$P<0.05$）；除了与$X3$、$X10$相关程度较弱，变量$X2$与其他变量间都呈显著相关性（$P<0.01$）；此外，变量$X3$与$X5$、$X9$；$X4$与$X5$、$X6$、$X7$；$X5$与$X7$；$X6$与$X7$、$X8$、

表1 研究区土壤有机质（SOM）含量的描述统计

Tab. 1 Descriptive statistics of soil organic matter content in the study area

指标/个	样点数/个	最小值/g·kg^{-1}	最大值/g·kg^{-1}	平均值/g·kg^{-1}	标准差/g·kg^{-1}	变异系数/%
整体	407	6.59	27.80	15.54	3.80	24.45
训练集	285	6.59	24.78	15.64	3.79	24.23
验证集	122	7.21	27.80	15.34	3.84	25.36

chinaXiv:202108.00019v1

ChinaXiv合作期刊

表2 变量相关分析

Tab. 2 Variable correlation analysis

	SOM	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
SOM	1.000										
X1	−0.315**	1.000									
X2	−0.099*	0.317**	1.000								
X3	−0.006	0.123*	0.082	1.000							
X4	−0.006	−0.073	0.166**	0.002	1.000						
X5	0.003	−0.016	−0.153**	−0.205**	−0.262**	1.000					
X6	−0.130**	0.336**	0.936**	0.058	0.174**	−0.074	1.000				
X7	0.101*	−0.340**	−0.475**	−0.053	0.222**	−0.181**	−0.458**	1.000			
X8	0.206**	−0.417**	−0.226**	−0.053	−0.057	0.099*	−0.229**	0.084	1.000		
X9	0.285**	−0.981**	−0.342**	−0.137**	0.021	0.038	−0.362**	0.306**	0.494**	1.000	
X10	0.204**	−0.235**	−0.089	0.004	−0.026	−0.016	−0.063	0.085	0.117*	0.185**	1.000

注:*表示相关性在0.05水平上显著,**表示相关性在0.01水平上显著(双尾);变量X1、X2、X3、X4、X5、X6、X7、X8、X9和X10分别代表高程、坡度、坡向、剖面曲率、平面曲率、地形起伏度、地形湿度指数、年均降水量、年均气温和归一化植被指数。下同。

X9;X7与X9;X8与X9;X9与X10间均在0.01置信水平上呈显著相关。相关性较强,说明变量之间存在较多的冗余信息,进一步表明对SOM预测建模的10个环境变量进行主成分提取是必要的。

利用Matlab软件分别实现对SOM预测因子集合数据的PCA、KPCA 2种方法的降维,同时提取累计贡献率大于85.00%的主成分,结果如表3所示。表3可以看出,PCA提取出能够反映原始变量90.00%信息量的6个主成分,基本可以概括原始变量所反映的信息,因此选择前6个主成分作为PCA组合模型的输入特征数据。由主成分载荷矩阵可知(表4):第一主成分与Slope(X2)、RDLS(X6)存在较大相关性,它反映了不同坡度、起伏条件下的SOM含量差异,可以概括为坡度因子;第二主成分与DEM(X1)、年均气温(X9)两因子具有较大相关性,反映了不同高程条件下的气温差异对SOM含量的影响,气温随着海拔的升高而逐渐降低,微生物的分解速率减慢致使SOM含量降低,概括为高程气温因子;第三主成分与NDVI(X10)、年均降水量(X8)存在较大相关性,概括为植被因子;第四、五主成分分别为剖面曲率、平面曲率因子;第六主成分与Aspect(X3)具有较大相关性,概括为坡向因子。不同坡向下的光照、温度、水分状况有差异,影响微生物活动与养分积累。相比之下,当核函数为多项式核函数时,KPCA可以提取出累计贡献率达到97.99%的2个主成分,综合考虑因子间相关性和主成分方差贡献率,认为这2个主成分包含了绝大部分的原始信息量,故选择前2个主成分作为KPCA组合模型的输入特征数据。

基于PCA、KPCA 2种主成分提取方法构建的SOM含量预测模型的输入特征数据,对RF、SVR、KNN模型分别进行训练,构建相对应的主成分和机器学习组合模型。不同参数的选取对模型的学习性能和预测精度有重要的影响,参数的过大过小都

表3 各主成分的贡献率

Tab. 3 Contribution ratio of each principal component factor

主成分	主成分分析(PCA)			核主成分分析(KPCA)	
	特征值	贡献率/%	累积贡献率/%	贡献率/%	累积贡献率/%
1	3.22	32.21	32.21	82.41	82.41
2	1.63	16.28	48.48	15.58	97.99
3	1.45	14.48	62.97	–	–
4	1.11	11.06	74.03	–	–
5	0.93	9.28	83.31	–	–
6	0.67	6.69	90.00	–	–

chinaXiv:202108.00019v1

ChinaXiv合作期刊

表4 主成分分析(PCA)载荷矩阵

Tab. 4 Principal component analysis load matrix

变量	主成分					
	1	2	3	4	5	6
*X*1	0.198	−0.952	0.001	−0.063	0.011	0.057
*X*2	0.935	−0.148	−0.113	0.113	−0.127	0.016
*X*3	0.031	−0.077	−0.019	−0.003	−0.106	0.987
*X*4	0.093	0.029	−0.067	0.953	−0.104	0.006
*X*5	−0.037	0.015	0.061	−0.122	0.972	−0.111
*X*6	0.925	−0.180	−0.074	0.149	−0.048	−0.006
*X*7	−0.672	0.187	−0.071	0.404	−0.247	−0.074
*X*8	−0.063	0.539	0.653	−0.007	0.072	0.043
*X*9	−0.203	0.962	0.055	0.009	0.005	−0.063
*X*10	−0.069	−0.075	0.919	−0.077	0.032	−0.041

会导致模型的过拟合或欠拟合问题。进一步遍历选择各模型的重要参数，根据10折交叉验证结果确定最优值(表5)。RF组合模型中确立的重要参数包括有决策树数量(ntree)和每棵决策树包括的特征数(mtry)；SVR组合模型在选择性能最优的radial核函数时，遍历选取初始化超参数松弛变量系数(*C*)和高斯核函数系数(gamma)的最优值。KNN组合模型建立过程中需要确立参数*K*的值：如果*K*值过小，异常的噪声点会对预测值产生较大的误差；过大的*K*值则会导致模型过于简单，学习的近似误差增大。

3.3 SOM空间分布预测结果

利用训练好的预测模型对研究区农耕区的SOM含量进行预测，生成咸阳市SOM空间分布图(图2)。由图2来看：(1)各模型预测结果整体上均呈现南高北低的分布特征，SOM含量的空间变化能够有效地反映地形地势信息，基本表现为南部平原区SOM含量高于北部高原、山地区，不同模型对SOM含量的空间变异趋势有较好的展示。(2) PCA组合模型预测结果高值区和低值区存在一定程度的“跳变”，尤其突出表现在山地区同一位置的山谷与山脊区域、平原区与高原区过渡区域；相比之下，KPCA组合模型预测结果空间上更为平滑，对于平原区SOM含量变化能够精细的呈现，更加符合实际情况。这种差异性很大程度上反映了PCA、KPCA 2种主成分提取方法的差异。PCA方法提取的主成分承载了过多易发掘的线性地形信息，致使模型预测结果与地形信息的过度拟合现象。(3) KPCA-SVR模型预测结果空间变异程度较大，难以反映SOM含量局部变异情况；KPCA-RF、KPCA-KNN模型对于展示SOM含量空间变化、预测SOM空间分布具有较好的预测效果。

表5 模型参数设定

Tab. 5 Model parameter setting

模型	参数设定
PCA-RF	ntree=1300、mtry=3
KPCA-RF	ntree=1300、mtry=2
PCA-SVR	radial核函数、*C*=1000、gramma=0.1
KPCA-SVR	radial核函数、*C*=1000、gramma=0.01
PCA-KNN	*K*=11
KPCA-KNN	*K*=3

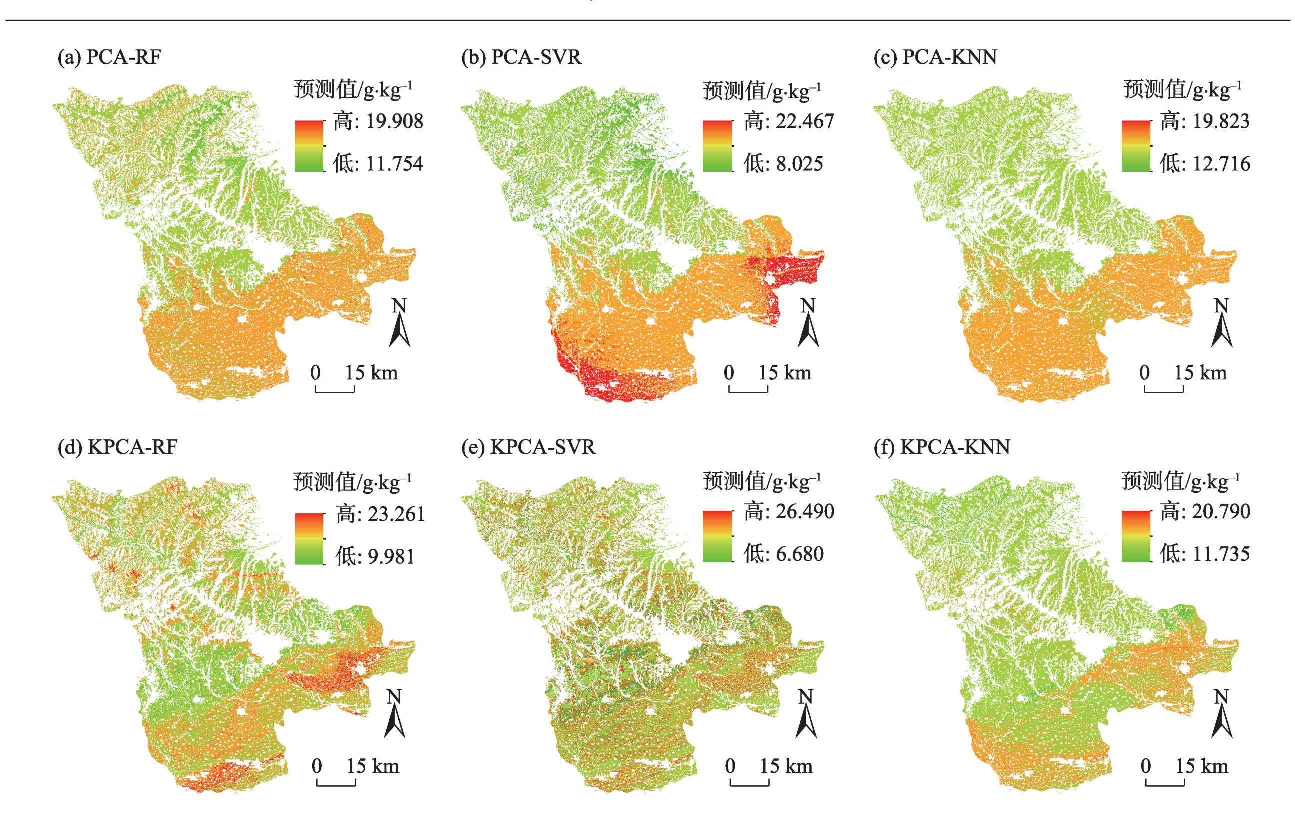
注：PCA-RF为主成分分析-随机森林组合模型；KPCA-RF为核主成分分析-随机森林组合模型；PCA-SVR为主成分分析-支持向量回归组合模型；KPCA-SVR为核主成分分析-支持向量回归组合模型；PCA-KNN为主成分分析-*K*最近邻组合模型；KPCA-KNN为核主成分分析-K最近邻组合模型。ntree为决策树数量；mtry为每棵决策树包括的特征数；*C*为初始化超参数松弛变量系数；gamma为高斯核函数系数；*K*为KNN模型参数。下同。

3.4 预测精度评价

为了进一步比较各组合模型的预测效果，选择未进行主成分分析的RF、SVR和KNN单一模型作为对比，对相同的数据集进行模型训练和验证，结果如表6所示。对比发现：PCA-RF组合模型预测精度较RF模型提高，决定系数R^2提高了0.023，RMSE、RAE误差值分别降低了0.070 g·kg^{-1}、2.440%；而PCA-SVR、PCA-KNN 2个预测模型较单一模型优化效果并不理想。KPCA-RF组合模型较RF模型精度提高，R^2、RMSE、RAE分别为0.791、1.970 g·kg^{-1}、50.100%，与单一RF模型相比，3个指标分别优化了

chinaXiv:202108.00019v1

ChinaXiv合作期刊



注:PCA-RF为主成分分析-随机森林组合模型;KPCA-RF为核主成分分析-随机森林组合模型;PCA-SVR为主成分分析-支持向量回归机组合模型;KPCA-SVR为核主成分分析-支持向量回归机组合模型;PCA-KNN为主成分分析-K最近邻组合模型;KPCA-KNN为核主成分分析-K最近邻组合模型。

图2 土壤有机质(SOM)含量空间分布预测图

Fig. 2 Prediction of spatial distribution of soil organic matter

表6 模型预测精度评价表

Tab. 6 Model accuracy verification table

模型	R^2	RMSE/g·kg^{-1}	RAE/%
PCA-RF	0.739	2.090	49.473
KPCA-RF	0.791	1.970	50.100
RF	0.716	2.160	51.913
PCA-SVR	0.479	2.936	66.718
KPCA-SVR	0.561	2.361	51.302
SVR	0.557	2.783	53.072
PCA-KNN	0.463	2.821	70.273
KPCA-KNN	0.507	2.676	66.079
KNN	0.490	2.744	68.553

注:RF为随机森林;SVR为支持向量回归机;KNN为K最近邻;R^2、RMSE和RAE分别表示决定系数、均方根误差和相对绝对误差。

0.075、0.190 g·kg^{-1}、1.813%;KPCA-SVR、KPCA-KNN组合模型预测精度较单一SVR、KNN模型提高。表明采用主成分提取方法和机器学习构建SOM含量预测组合模型能够消除数据间冗余和相关性,在一定程度上可以提高模型预测精度。同时发现:KPCA组合模型的SOM含量预测模型精度高于相应的PCA组合模型,这充分展示了KPCA在降低原始变量维度,处理数据间非线性关系的优势性能。综合来看,KPCA-RF模型对SOM含量预测精度高于其他模型,该模型对于咸阳市农耕区SOM含量具有较好的预测效果,可以为SOM空间预测与数字制图提供科学依据。

4 讨论

机器学习方法在土壤养分空间预测中具有巨大的潜力,主成分提取方法可以有效解决高变量维度小样本量的数据分析问题。已有研究表明,组合主成分提取方法和机器学习算法可以有效提升模型精度,优化模型性能[37-38]。本文利用PCA线性提取方法、KPCA非线性提取方法与RF、SVR、KNN模型组合构成6种组合预测模型,使用10折交叉验证,比较了不同组合模型对SOM含量预测性能。在本研究中,使用较少的主成分即达到较为满意的预测精度,主成分提取方法有效降低了数据相关性,

chinaXiv:202108.00019v1

ChinaXiv合作期刊

提升了模型精度。对比发现，KPCA方法考虑了数据间的非线性关系，相应的组合模型预测精度优于PCA组合模型。在本文研究的6种组合模型中，KPCA-RF预测精度最高，可以应用于SOM空间预测制图，如许杏花等[39]对风电场功率预测的研究也证明了KPCA-RF具有较高的准确性。但是值得关注的是，KPCA方法自身也存在着一些问题，主要是核函数非线性映射后的矩阵很难用物理意义解释。其次，KPCA计算量只决定于数据集的大小，样本点较大时带来的核矩阵维数较大，造成计算量增加，建议从样本集中选择有代表性的学习样本来降低核矩阵阶数，已有的方法包括疏散的核PCA方法[40]、模糊C-均值聚类遗传算法[41]等。

基于最优模型KPCA-RF模型预测得到咸阳市SOM含量均值为15.56±1.96 g·kg^{-1}，整体上与样点统计结果保持一致。SOM空间分布呈现南部较高、北部较低，这与任丽[42]、赵叶婷[43]等在该地区的研究结果一致，这种差异主要是受地形条件影响造成的。南部为关中平原区，海拔较低且地形平坦，北部海拔较高，地形条件复杂。关中平原区长久以来稳定的农业生产保证了SOM一直处于相对较高的水平，相比之下，北部地区复杂的地形条件和气候条件对农业生产造成了很大的困难。此外，随着海拔的升高，SOM存在明显的降低趋势，这主要是因为从高海拔表土冲刷而来的SOM在低海拔沉积，同时受水蚀作用影响，土壤蓄水保肥能力降低[44]。

信息科学与数据科学的发展使得人们获取精准农业信息成为可能，协同多源环境变量的诸多机器学习算法在揭示SOM空间变异与空间制图方面得到了广泛的应用。SOM空间变异是自然环境与人类活动综合作用结果，自然环境因子易于空间化而被广泛应作建模指标，人类活动因子因不具备空间连续性难以参与模型训练，进一步的研究方向应探讨解决人为因子空间化的问题。通过对采样点数据不同人类管理措施下的SOM含量统计发现，不同种植制度、灌溉条件下SOM含量差异明显，表明了人为管理措施是SOM空间变异研究中不可忽略的方面[42]。特别是在县、乡等更小尺度上的研究中，人类耕作方式与强度对SOM含量的差异影响较大[45-46]，仅依靠自然环境变量很难实现精准预测。因此，人为因子的采集与空间量化对于提升模型精度有重要的意义。

5 结 论

本文提出和构建了基于主成分和机器学习的土壤养分含量空间预测模型，并在咸阳市农耕区SOM含量的空间预测中取得了较好的应用效果。主要结论如下：

（1）使用PCA、KPCA 2种主成分提取方法，实现了数据降维，消除了变量间的相关性和冗余性，有利于提升土壤养分含量预测模型的精度和稳定性。

（2）基于KPCA特征提取和机器学习算法的组合预测模型与单一预测模型、PCA组合模型相比，预测精度较高，能够很好的拟合土壤养分含量与环境变量之间的非线性关系。

（3）构建的KPCA-RF组合模型与KPCA-SVR等其他模型相比，模型评价指标决定系数、均方根误差和相对绝对误差分别为0.791、1.970 g·kg^{-1}和50.100%，优于其他预测模型。该模型对咸阳市农耕区SOM含量的预测取得了良好的效果，可以进一步运用到其余土壤养分的精准预测与地力评价中。

参考文献（References）

[1] 王绍强, 周成虎, 李克让, 等. 中国土壤有机碳库及空间分布特征分析[J]. 地理学报, 2000, 67(5): 533–544. [Wang Shaoqiang, Zhou Chenghu, Li Kerang, et al. Analysis on spatial distribution characteristics of soil organic carbon reservoir in China[J]. Acta Geographica Sinica, 2000, 67(5): 533–544.]

[2] 朱阿兴, 杨琳, 樊乃卿, 等. 数字土壤制图研究综述与展望[J]. 地理科学进展, 2018, 37(1): 66–78. [Zhu A'xing, Yang Lin, Fan Naiqing, et al. The review and outlook of digital soil mapping[J]. Progress in Geography, 2018, 37(1): 66–78.]

[3] Scull P, Franklin J, Chadwick O A, et al. Predictive soil mapping: A review[J]. Progress in Physical Geography, 2003, 27(2): 171–197.

[4] Moore I D, Gessler P E, Nielsen G A E, et al. Soil attribute prediction using terrain analysis[J]. Soil Science Society of America Journal, 1993, 57(2): 443–452.

[5] 张甘霖, 朱阿兴, 史舟, 等. 土壤地理学的进展与展望[J]. 地理科学进展, 2018, 37(1): 57–65. [Zhang Ganlin, Zhu A'xing, Shi Zhou, et al. Progress and future prospect of soil geography[J]. Progress in Geography, 2018, 37(1): 57–65.]

[6] Jahan N, Gan T Y. Modelling the vegetation-climate relationship in a boreal mixedwood forest of Alberta using normalized difference and enhanced vegetation indices[J]. International Journal of

chinaXiv:202108.00019v1

ChinaXiv合作期刊

Remote Sensing, 2011, 32(2): 313-335.

[7] Thompson J A, Pena Yewtukhiw E M, Grove J H. Soil-landscape modeling across a physiographic region: Topographic patterns and model transportability[J]. Geoderma, 2006, 133(1-2): 57-70.

[8] Zhu A X, Band L, Vertessy R, et al. Derivation of soil properties using a soil land inference model (SoLIM)[J]. Soilence Society of America Journal, 1997, 61(2): 523-533.

[9] 郭旭东, 傅伯杰, 马克明, 等. 基于GIS和地统计学的土壤养分空间变异特征研究——以河北省遵化市为例[J]. 应用生态学报, 2000(4): 557-563. [Guo Xudong, Fu Bojie, Ma Keming, et al. Spatial variability of soil nutrients based on geostatistics combined with GIS: A case study in Zunhua City of Hebei Province[J]. Chinese Journal of Applied Ecology, 2000(4): 557-563.]

[10] Frogbrook Z L, Oliver M A. Comparing the spatial predictions of soil organic matter determined by two laboratory methods[J]. Soil Use and Management, 2001, 17(4): 235-244.

[11] Poggio T, Smale S. The mathematics of learning: Dealing with data [J]. Notices of the American Mathematical Society, 2003, 50(5): 537-544.

[12] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[13] 王茵茵, 齐雁冰, 陈洋, 等. 基于多分辨率遥感数据与随机森林算法的土壤有机质预测研究[J]. 土壤学报, 2016, 53(2): 342-354. [Wang Yinyin, Qi Yanbing, Chen Yang, et al. Prediction of soil organic matter based on multi-resolution remote sensing data and random forest algorithm[J]. Acta Pedologica Sinica, 2016, 53 (2): 342-354.]

[14] 郭澎涛, 李茂芬, 罗微, 等. 基于多源环境变量和随机森林的橡胶园土壤全氮含量预测[J]. 农业工程学报, 2015, 31(5): 194-200. [Guo Pengtao, Li Maofen, Luo Wei, et al. Prediction of soil total nitrogen for rubber plantation at regional scale based on environmental variables and random forest approach[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(5): 194-200.]

[15] Lu Y Y, Liu F, Zhao Y G, et al. An integrated method of selecting environmental covariates for predictive soil depth mapping[J]. Journal of Integrative Agriculture, 2019, 18(2): 301-315.

[16] Tomislav H, Heuvelink G B M, Bas K, et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions[J]. Plos One, 2015, 10(6): e0125814, doi: 10.1371/journal.pone.0125814.

[17] Malone B P, McBratney A B, Minasny B, et al. Mapping continuous depth functions of soil carbon storage and available water capacity[J]. Geoderma, 2009, 154(1-2): 138-152.

[18] Mansuy N, Thiffault E, Paré D, et al. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the K-nearest neighbor method[J]. Geoderma, 2014, 235: 59-73.

[19] Curtis J R, Newell R K, James J C, et al. Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations[J]. Computers and Electronics in Agriculture, 2019, 164: 104872, doi: 10.1016/j.compag.2019.104872.

[20] 任丽, 杨联安, 王辉, 等. 基于随机森林的苹果区土壤有机质空间预测[J]. 干旱区资源与环境, 2018, 32(8): 141-146. [Ren Li, Yang Lian'an, Wang Hui, et al. Spatial prediction of soil organic matter in apple region based on random forest[J]. Journal of Arid Land Resources and Environment, 2018, 32(8): 141-146.]

[21] Forkuor G, Hounkpatin O K L, Welp G, et al. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models[J]. Plos One, 2017, 12(1): e0170478, doi: 10.1371/journal.pone.0170478.

[22] 袁玉琦, 陈瀚阅, 张黎明, 等. 基于多变量与RF算法的耕地土壤有机碳空间预测研究——以福建亚热带复杂地貌区为例[J/OL]. [2020-12-21]. https://kns.cnki.net/kcms/detail/32.1119.P.20200824.1432.002.html. [Yuan Yuqi, Chen Hanyue, Zhang Liming, et al. Prediction of spatial distribution of soil organic carbon in farmland based on multi-variables and random forest algorithm: A case study of a subtropical complex geomorphic region in Fujian as an example[J/OL]. [2020-12-21]. http://kns.cnki.net/kcms/detail/32.1119.P.20200824.1432.002.html.]

[23] 张振华, 丁建丽, 王敬哲, 等. 集成土壤-环境关系与机器学习的干旱区土壤属性数字制图[J]. 中国农业科学, 2020, 53(3): 563-573. [Zhang Zhenhua, Ding Jianli, Wang Jingzhe, et al. Digital soil properties mapping by ensembling soil-environment relationship and machine learning in arid regions[J]. Scientia Agricultura Sinica, 2020, 53(3): 563-573.]

[24] 王念一, 于丰华, 许童羽, 等. 基于机器学习的粳稻叶片叶绿素含量高光谱反演建模[J]. 浙江农业学报, 2020, 32(2): 359-366. [Wang Nianyi, Yu Fenghua, Xu Tongyu, et al. Hyperspectral retrieval modelling for chlorophyll contents of japonica-rice leaves based on machine learning[J]. Acta Agriculturae Zhejiangensis, 2020, 32(2): 359-366.]

[25] Zheng L, Watson D G, Johnston B F, et al. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with PCA, support vector machines and random forest data modeling[J]. Analytica Chimica Acta, 2009, 642(1-2): 257-265.

[26] 聂红梅, 杨联安, 李新尧, 等. 基于PCA-SVR的冬小麦土壤水分预测[J]. 土壤, 2018, 50(4): 812-818. [Nie Hongmei, Yang Lian'an, Li Xinyao, et al. Prediction of soil moisture of winter wheat by PCA-SVR[J]. Soils, 2018, 50(4): 812-818.]

[27] 刘新华, 杨勤科, 汤国安. 中国地形起伏度的提取及在水土流失定量评价中的应用[J]. 水土保持通报, 2001, 21(1): 57-59, 62. [Liu Xinhua, Yang Qinke, Tang Guo'an. Extraction and application of relief of China based on DEM and GIS method[J]. Bulletin of Soil and Water Conservation, 2001, 21(1): 57-59, 62.]

[28] 张彩霞, 杨勤科, 李锐. 基于DEM的地形湿度指数及其应用研究进展[J]. 地理科学进展, 2005, 24(6): 116-123. [Zhang Caixia,

chinaXiv:202108.00019v1

ChinaXiv合作期刊

Yang Qinke, Li Rui. Advancement in topographic wetness index and its application[J]. Progress in Geography, 2005, 24(6): 116-123.]

[29] 李朝荣, 刘扬, 李春明. PCA与KPCA在综合评价中的应用[J]. 宜宾学院学报, 2010, 10(12): 27-30. [Li Chaorong, Liu Yang, Li Chunming. Application of PCA and KPCA in comprehensive evaluation[J]. Journal of Yibin University, 2010, 10(12): 27-30.]

[30] 王瀛, 郭雷, 梁楠. 基于优选样本的KPCA高光谱图像降维方法[J]. 光子学报, 2011, 40(6): 847-851. [Wang Ying, Guo Lei, Liang Nan. A dimensionality reduction method based on KPCA with optimized sample set for hyperspectral image[J]. Acta Photonica Sinica, 2011, 40(6): 847-851.]

[31] 杨道军, 钱新, 钱瑜, 等. 核主成分分析法在生态经济可持续发展评价中应用[J]. 环境科学与技术, 2007(12): 91-93, 122. [Yang Daojun, Qian Xin, Qian Yu, et al. Application of kernel principal component analysis in evaluation of sustainable development of ecological economy[J]. Environmental Science & Technology, 2007(12): 91-93, 122.]

[32] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[33] Cutler A, Cutler D R, Stevens J R. Random forests[J]. Machine Learning, 2011, 45(1): 157-176.

[34] Awad M, Khanna R. Support vector regression[J]. Neural Information Processing Letters & Reviews, 2007, 11(10): 203-224.

[35] 毋雪雁, 王水花, 张煜东. K最近邻算法理论与应用综述[J]. 计算机工程与应用, 2017, 53(21): 1-7. [Wu Xueyan, Wang Shuihua, Zhang Yudong. Survey on theory and application of K-nearest- neighbors algorithm[J]. Computer Engineering and Applications, 2017, 53(21): 1-7.]

[36] 毕达天, 邱长波, 张晗. 数据降维技术研究现状及其进展[J]. 情报理论与实践, 2013, 36(2): 125-128. [Bi Datian, Qiu Changbo, Zhang Han. Current situation and latest development of research on data dimension reduction technology[J]. Information Studies: Theory & Application, 2013, 36(2): 125-128.]

[37] 刘炳春, 符川川, 李健. 基于PCA-SVR模型的中国CO_2排放量预测研究[J]. 干旱区资源与环境, 2018, 32(4): 56-61. [Liu Bingchun, Fu Chuanchuan, Li Jian. Forecast of CO_2 emission in China based on PCA-SVR[J]. Journal of Arid Land Resources and Environment, 2018, 32(4): 56-61.]

[38] 赵帅, 黄亦翔, 王浩任, 等. 基于随机森林与主成分分析的刀具磨损评估[J]. 机械工程学报, 2017, 53(21): 181-189. [Zhao Shuai, Huang Yixiang, Wang Haoren, et al. Random forest and principle components analysis based on health assessment methodology for tool wear[J]. Journal of Mechanical Engineering, 2017, 53(21): 181-189.]

[39] 许杏花, 潘庭龙. 基于KPCA-RF的风电场功率预测方法研究[J]. 可再生能源, 2018, 36(9): 1323-1327. [Xu Xinghua, Pan Tinglong. Wind power prediction based on KPCA-RF[J]. Renewable Energy Resources, 2018, 36(9): 1323-1327.]

[40] Michael E T, Cambridge C N. Sparse kernel principal component analysis[J]. Advances in Neural Information Processing Systems, 2001, 13: 633-639.

[41] 高新波, 谢维信. 模糊聚类理论发展及应用的研究进展[J]. 科学通报, 1999, 44(21): 3-5. [Gao Xinbo, Xie Weixin. Research progress on the development and application of fuzzy clustering theory [J]. Chinese Science Bulletin, 1999, 44(21): 3-5.]

[42] 任丽. 基于多源环境变量的土壤养分预测及综合评价[D]. 西安: 西北大学, 2019. [Ren Li. Spatial prediction and comprehensive evaluation of soil nutrients based on environmental variables [D]. Xi'an: Northwest University, 2019.]

[43] 赵业婷. 基于GIS的陕西省关中地区耕地土壤养分空间特征及其变化研究[D]. 杨凌: 西北农林科技大学, 2015. [Zhao Yeting. Spatial characteristics and changes of soil nutrients in cultivated land of Guanzhong region in Shaanxi Province based on GIS[D]. Yangling: Northwest A & F University, 2015.]

[44] 邱扬, 傅伯杰, 王军, 等. 黄土高原小流域土壤养分的时空变异及其影响因子[J]. 自然科学进展, 2004, 14(3): 56-61. [Qiu Yang, Fu Bojie, Wang Jun, et al. Temporal and spatial variability and influencing factors of soil nutrients in small watersheds of the Loess Plateau[J]. Progress in Natural Science, 2004, 14(3): 56-61.]

[45] 杨景成, 韩兴国, 黄建辉, 等. 土壤有机质对农田管理措施的动态响应[J]. 生态学报, 2003, 23(4): 787-796. [Yang Jingcheng, Han Xingguo, Huang Jianhui, et al. The dynamics of soil organic matter in cropland responding to agricultural practices[J]. Acta Ecologica Sinica, 2003, 23(4): 787-796.]

[46] 宋明伟, 李爱宗, 蔡立群, 等. 耕作方式对土壤有机碳库的影响[J]. 农业环境科学学报, 2009, 27(2): 224-228. [Song Mingwei, Li Aizong, Cai Liqun, et al. Effects of different tillage methods on soil organic carbon pool[J]. Journal of Agro-Environment Science, 2009, 27(2): 224-228.]

chinaXiv:202108.00019v1

ChinaXiv合作期刊

Spatial prediction modeling of soil organic matter content based on principal components and machine learning

HU Guigui[1,2], YANG Fenli[3], YANG Lian'an[1,2], ZHENG Yurong[1,2],
WANG Hui[4], CHEN Weijun[5], LI Yali[1,2]

(1. Shaanxi Key Laboratory of Earth Surface System and Environmental Carrying Capacity, Northwest University, Xi'an 710127, Shaanxi, China; 2. College of Urban and Environmental Sciences, Northwest University, Xi'an 710127, Shaanxi, China; 3. Xianyang Station of Soil and Fertilizer, Xianyang 712000, Shaanxi, China; 4. Academy of Agriculture Sciences of Xianyang, Xianyang 712000, Shaanxi, China; 5. Xunyi Station of Soil and Fertilizer, Xunyi 711300, Shaanxi, China)

Abstract: Spatial prediction models of soil nutrients are constructed from collaborative environment variables and machine learning regression models; they are of great significance for accurate nutrient management, but the information redundancy and correlation among multidimensional variables can lead to problems such as a long training time for the model and low prediction accuracy. In this study, the farming area of Xianyang City, Shaanxi Province, China, was taken as an example, and 10 environmental variables were selected: the elevation, aspect, slope, plane curvature, section curvature, relief, topographic wetness index, annual average temperature, annual average precipitation, and normalized difference vegetation index. Features were extracted by principal component analysis (PCA) and kernel PCA (KPCA), which were combined with the random forest (RF), support vector regression (SVR), and K nearest neighbor (KNN) models to develop spatial prediction models for the soil organic matter (SOM). Single models were used as the control. Then, the prediction accuracy of different models was evaluated according to the model determination coefficient (R^2), root-mean-squared error (RMSE), and relative absolute error (RAE). The following results were obtained: (1) PCA and KPCA reduced the data dimensionality, which eliminated the correlation and redundancy between variables and helped improve the accuracy and stability of the SOM spatial prediction model. (2) The PCA-RF model had a higher prediction accuracy than the RF model (R^2 increased by 0.023, RMSE and RAE decreased by 0.070 $g \cdot kg^{-1}$ and 2.440%, respectively), whereas PCA-SVR and PCA-KNN performed worse than SVR and KNN alone. (3) The KPCA-RF model had higher accuracy than the RF model (R^2, RMSE, and RAE were 0.791, 1.970 $g \cdot kg^{-1}$, and 50.100%, respectively). The KPCA-SVR and KPCA-KNN models had better prediction accuracies than the SVR and KNN models. (4) The combined prediction model based on KPCA feature extraction and machine learning had higher prediction accuracy than the PCA-based combined prediction models and single prediction models and fitted well to the nonlinear relationship between the SOM content and environmental variables. The KPCA-RF model performed better than the other prediction models. This model accurately predicted the SOM content in the agricultural area of Xianyang City, and it can be further applied to accurately predicting other soil nutrients and evaluating soil fertility.

Key words: soil organic matter; machine learning; kernel principal component analysis; farming area; Xianyang City

chinaXiv:202108.00019v1